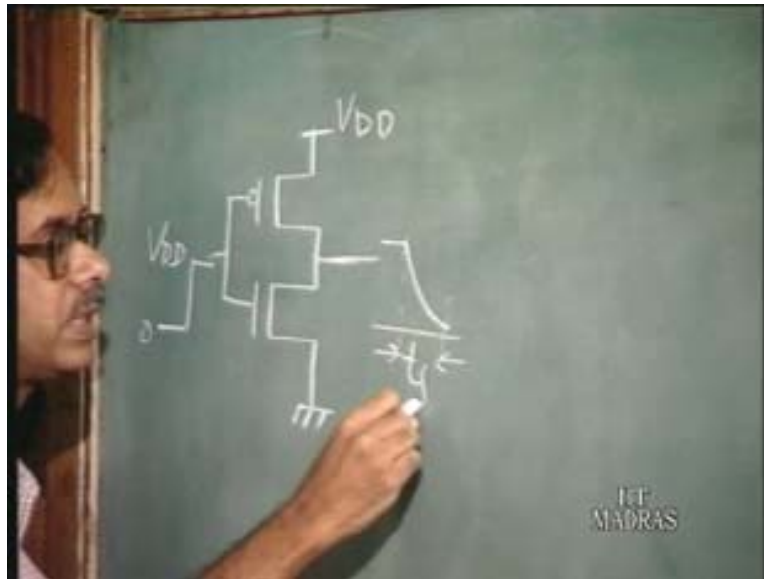


Digital Integrated Circuits
Dr. Amitava Dasgupta
Department of Electrical Engineering
Indian Institute of Technology, Madras
Lecture No # 27

CMOS NAND, NOR and the other gates:
Clocked CMOS

We shall continue our discussion on CMOS logic circuits. We were discussing the delays during switching of a CMOS inverter and we are discussing this case when the input voltage of the CMOS inverter is switched from low to high say 0 to V_{DD} and we said that the output voltage is going to go from high to low. What is the time required? We call it the fall time and we had also seen the relation for this fall time and this fall time can be evaluated by integrating the current charging the capacitance and we get a relation like this $t_{f,fall} = \frac{C_L}{K_n} \ln \left(\frac{V_{OH} - V_{in}}{V_{OL} - V_{in}} \right)$ by $V_{in} - V_T$ plus half \ln twice $V_{in} - V_T - V_{OL}$ divided by V_{OL} .

(Refer Slide Time: 01:45)



That is the time required for the output to discharge from the output high or V_{OH} to a low of V_{OL} so that is the time required. We have already seen this and we see that this fall time $t_{ow,f}$ is proportional to C_L by K_n . We have also said that this C_L which is the load capacitance is the input capacitance of a similar gate, if this particular inverter is driving a similar gate so the C_L is the input capacitance of a gate in that particular situation. We had said that the input capacitance of a gate, we shall assume although it consists of different capacitance components we shall assume to be, mostly due to the oxide capacitance and it is given by C_{ox} which is the oxide capacitance per unit area

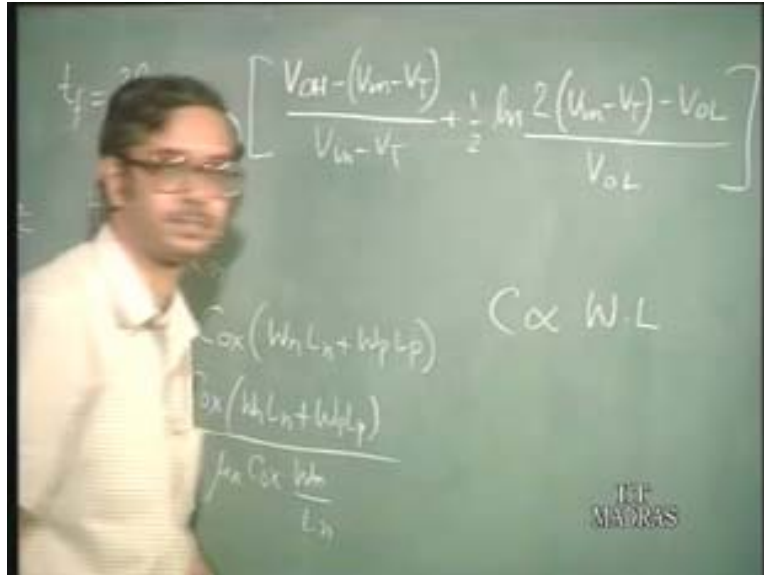
into $W_n L_n$ plus $W_p L_p$ which is the total input area of the gate area, $W_n L_n$ corresponds to the NMOS and $W_p L_p$ corresponds to the PMOS.

(Refer Slide Time: 05:39)



Now if this is the situation this C_L by K_n is equal to C_{ox} $W_n L_n$ plus $W_p L_p$ divided by $\mu_n C_{ox} W_n$ by L_n . We have also seen that because of the mobility difference between the holes and electrons, the sizes of the n channel and p channel MOSFET must be made different, if you want to get a symmetric input output characteristics. Obviously one of this device must have a dimension which is bigger than the other. Now which dimension do we make small and which dimension do we make big? Now if you look at this C_L by K_n , C is proportional to the capacitance is proportional to the area. Area means C is proportional to W into L and K that is the trans conductor is proportional to W by L .

(Refer Slide Time: 05:58)



So C by K would depend on L squared. We would like to make this C by K as small as possible to get a shorter time of discharge of the capacitance.

(Refer Slide Time: 06:44)

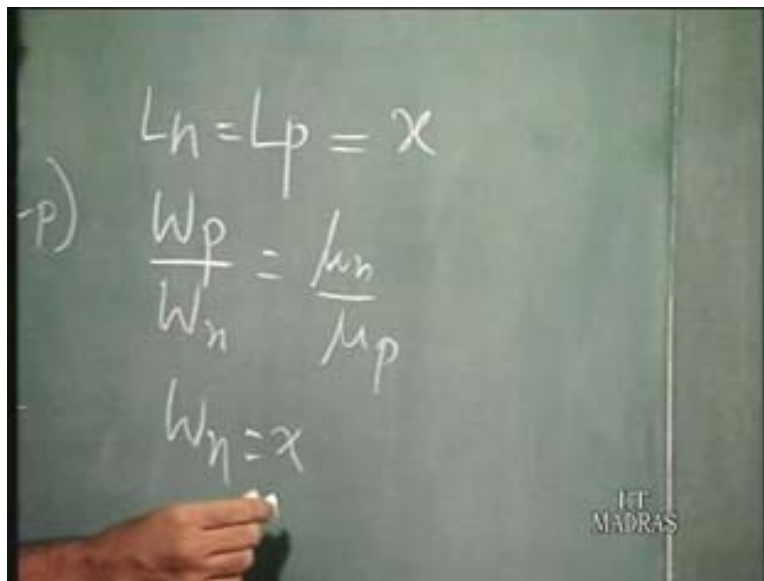


This C by K in order to make it small, this is proportional to L square. So the length should be made as small as possible. You see that the width, it is not so much dependent on width. Why? Because if you are increasing the width, you are increasing the capacitance on one hand but you are also increasing the current which is driving that capacitance. So these two effects cancel out and the discharge time remains constant more or less constant. The length has to be minimized. So what we do is in this situation, we make L_n is equal to L_p as the minimum size of the device or as the

minimum possible size which you can fabricate or the column minimum feature size in a particular circuit.

So obviously L_n is equal to L_p and what about the widths, W_p has to be larger than W_n . We have seen that W_p by W_n is going to be equal to μ_n by μ_p . So obviously what we can do is W_n we can also make as the minimum feature size and W_p becomes μ_n by μ_p times the minimum feature size. So W_n is equal to x and W_p is equal to μ_n by μ_p x .

(Refer Slide Time: 08:09)



A photograph of a chalkboard with handwritten equations. The equations are:

$$L_n = L_p = x$$
$$\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p}$$
$$W_n = x$$

A hand is visible at the bottom left, holding a piece of chalk. In the bottom right corner, there is a small logo that reads "IIT MADRAS".

If we substitute these values here, back to C_L by K_n expression we see that C_{ox} goes, so W_n and L_n are both x in this. So this is x squared plus W_p . L_p , W_p is equal to μ_n by μ_p x and L_p is equal to x .

(Refer Slide Time: 08:17)

(P)

$$\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p}$$

$$W_n = x$$

$$W_p = \frac{\mu_n}{\mu_p} x$$

ITT
MADRAS

So you have x squared into μ_n by μ_p divided by, W_n by L_n is equal to 1, both of them are the minimum feature size so this becomes equal to μ_n , so this becomes equal to x squared 1 by μ_n plus 1 by μ_p . So that is the value of C_L by K_n .

(Refer Slide Time: 09:09)

$$\frac{C_L}{K_n} = \frac{C_{ox}(W_n L_n + W_p L_p)}{\mu_n C_{ox} W_n}$$

$$= \frac{x^2 + x^2 \frac{\mu_n}{\mu_p}}{\mu_n} = x^2 \left(\frac{1}{\mu_n} + \frac{1}{\mu_p} \right)$$

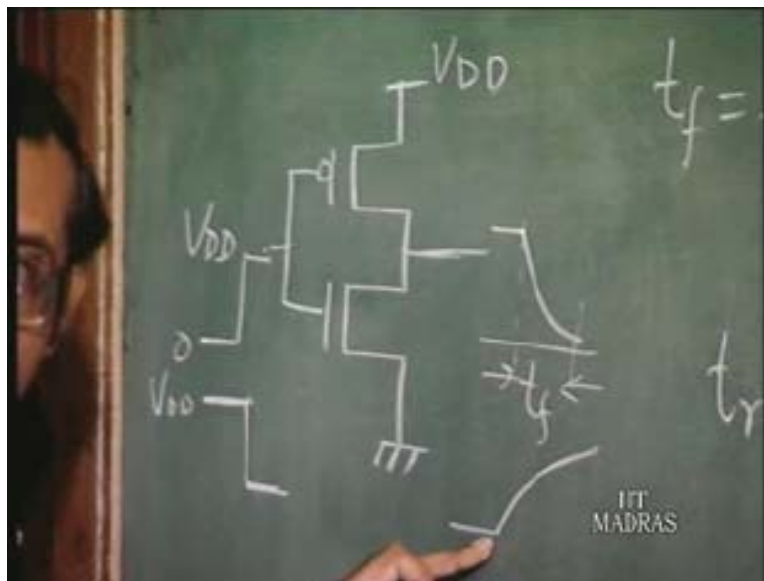
ITT
MADRAS

The important thing here is C_L by K_n is proportional to x square which is the minimum feature size in a particular technology with which you are fabricating this CMOS circuit. It is quite obvious that if you go on reducing the minimum feature size, the delays are going to come down. So that is one of the reasons why you see the dimensions nowadays are becoming smaller and smaller. The drive for that is to

achieve lower delays as also of course to have higher packing density. So the delay would be proportional to the square of the lengths of the devices which is the minimum feature size. This is for the fall time. If you consider the raise time that is when the input voltage goes from high to low, say if a sudden transmission from high to low, the output is going to raise from low to high.

So you would have a similar sort of analysis which one can do where the rise time will be equal to, suppose the output voltage is rising from zero to V_{OH} , so C_L by I_{dV} . Again we have to break it up into two regions of operation, the input voltage is say V_{DD} . If the input voltage is V_{DD} then output voltage initially is zero, it goes up and so initially it is in the linear region of operation. So when it goes up to V_{DD} minus V_{T} , it goes into the saturation region. So you have to break it up 0 to V_{DD} minus V_{T} in that case.

(Refer Slide Time: 11:10)



So C_L by I linear dV plus V_{DD} minus V_T to V_{OH} C_L by I_{sat} dV , so you do this and again you come up with an expression. I just write down the expression, final value twice C_L . What is happening is the input voltage is going from V_{DD} to 0, basically this NMOS is off. The PMOS is on, so it is this PMOS which is going to charge the capacitance. When this PMOS is charging this capacitance, the output voltage is raising from 0 to V_{DD} . Initially when the input voltage is equal to 0, the drain to source voltage of the PMOS is equal to the modulus of that is V_{DD} .

So initially this PMOS is in saturation, the gate to source voltage is also equal to minus of V_{DD} , so when this goes up to V_T . then what happens is this goes from the saturation to the linear region and so the PMOS is on, so the output voltage is going from zero to V_T C_L by I_{sat} dV plus V_T to V_{OH} C_L I_{lin} dv .

(Refer Slide Time: 14:49)

$$\begin{aligned}
 t_r &= \int_0^{V_{OH}} \frac{C_L}{I} dV \\
 &= \int_0^{V_T} \frac{C_L}{I_{sat}} dV + \int_{V_T}^{V_{OH}} \frac{C_L}{I_{lin}} dV \\
 &= \frac{2C_L}{K_p(V_{DD}-V_T)} \left[\frac{V_T}{V_{DD}-V_T} + \frac{1}{2} \ln \frac{V_{DD}-2V_T+V_{OH}}{V_{DD}-V_{OH}} \right]
 \end{aligned}$$

If you do this analysis, this is the PMOS which is conducting. So you get twice C_L by K_p . V_{DD} minus V_T . [V_T by V_{DD} minus V_T plus half \ln V_{DD} minus twice V_T plus V_{OH} by V_{DD} minus V_{OH}]. So this is the raise time expression which can be done by substituting the values of I_{sat} and I linear for the PMOS device. So you get C_L by K_p here and C_L by K_p is if you assume that C_L is the load capacitance, is again the input capacitance of a MOSFET, C_L by K_p will give you the similar relation as we have got for the other case that is equal to x squared 1 by μ_n plus 1 by μ_p .

(Refer Slide Time: 15:25)

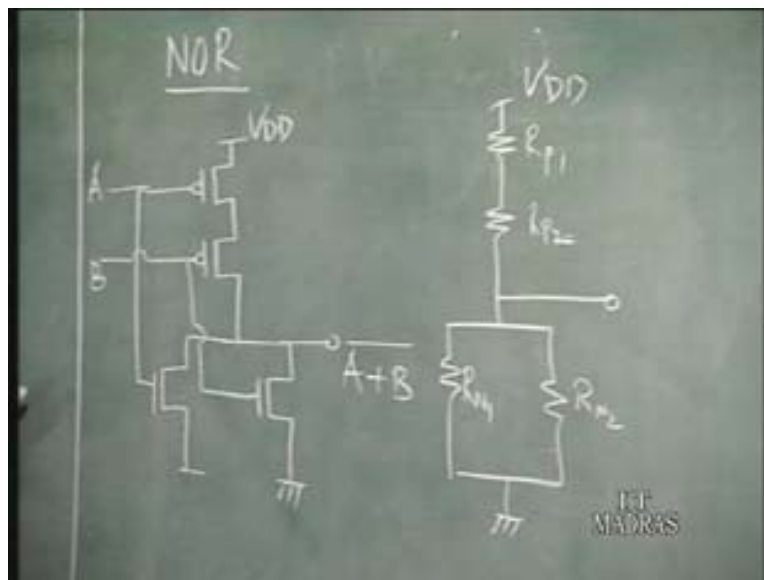
$$\frac{C_L}{K_p} = x^2 \left(\frac{1}{\mu_n} + \frac{1}{\mu_p} \right)$$

This delay is proportional to x squared again and you have the other terms here which is dependent on the different voltages. So by this way you can calculate the propagation

delay of the MOSFET of the CMOS device and we have seen that the delays are proportional to the square of the channel length. So up to now we have been discussing the inverter which consists of a PMOS and an NMOS device.

Now we would like to have the more complicated logic circuits basically for example the NAND gate and the NOR gates with which you can fabricate other logic gates, realize any logic function. So how do you make a NOR gate say for example. the NOR gate is fabricated in this way that is you have two NMOS devices in parallel and two PMOS devices in series and the gates of these two are shorted and these are the two inputs, this is V_{DD} and this is the output, A B and this is the NOR of A and B at the output. So how is this a NOR gate? We can just explain it in terms of the model which we had taken up that is for the resistances. Suppose we write it like this. so this is the output, you have V_{DD} . so this is R_{p1} . say, this is R_{p2} and this is say R_{n1} . and this is R_{n2} where R_n is the resistance of the n channel devices and R_p are the resistance of the p channel devices.

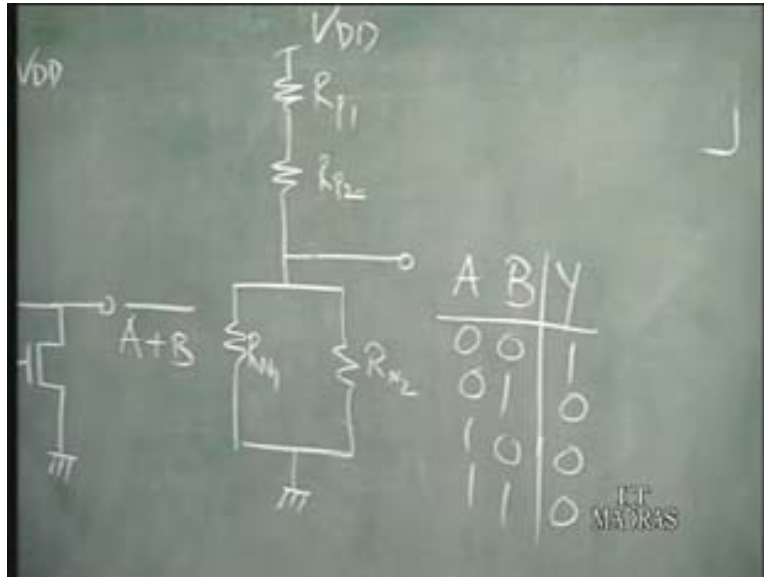
(Refer Slide Time: 18:03)



You must realize that if any input is low, what it does is it cuts off the n channel device and the corresponding p channel device is conducting. If any input is high, the n channel device is conducting and the corresponding P channel device is off. If I just put a truth table, y is the output so you can have the four combinations. If both the inputs are zeros what happens to the two NMOS devices? They are both off and the PMOS devices they are both on. The output is equal to V_{DD} which is high. If one input is high and the other input is low what happens is in the lower branch here one of the NMOS device is on and the other is off but they are in parallel.

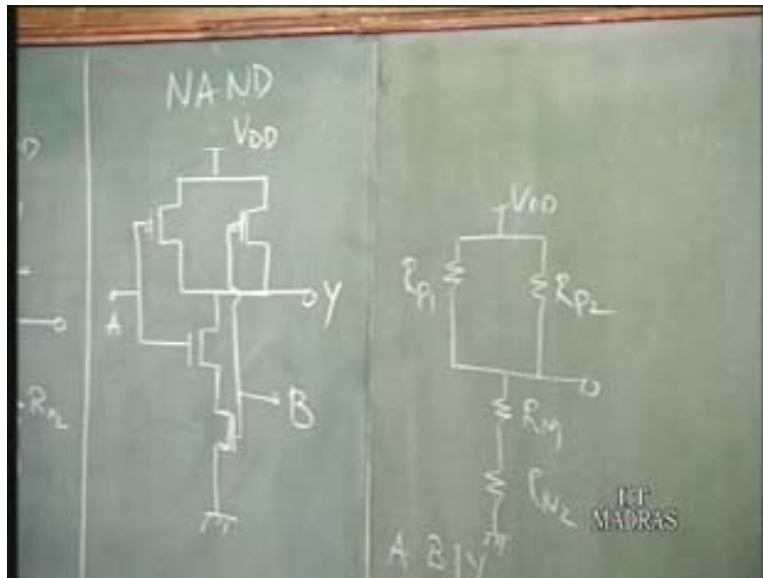
So the lower branch is conducting whereas here one of the device is on and the other is off but they are in series, so this upper branch is off, this is not conducting. So you have a path from the output to the ground which means that the output is going to be low. the output capacitance can be discharged and if both the inputs are high which means that both the NMOS devices are on and both the PMOS devices are off which means that the output is going to be again low so this is a NOR gate.

(Refer Slide Time: 19:40)



So in the NOR gate the two PMOS transistors in series and the two NMOS transistors in parallel. Now the NAND gate; in a NAND gate what we do is we have the 2 PMOS transistors in parallel and two NMOS transistors in series.

(Refer Slide Time: 23:10)



These are the two inputs say this input is A and suppose this input is B and this is the output. Again we can have a similar model of the NMOS N channel transistors, the resistance is being in series and for the PMOS transistors, the resistance being in parallel. So this R_{n1} , this is R_{n2} , R_{p1} and R_{p2} . Now what happens again if any input is high, it is basically going to switch on the NMOS device and cut off the PMOS device and if any input is low, the PMOS device is on and the NMOS device is off.

(Refer Slide Time: 22:41)

A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

A small logo 'IIT MADRAS' is visible in the bottom right corner of the table.

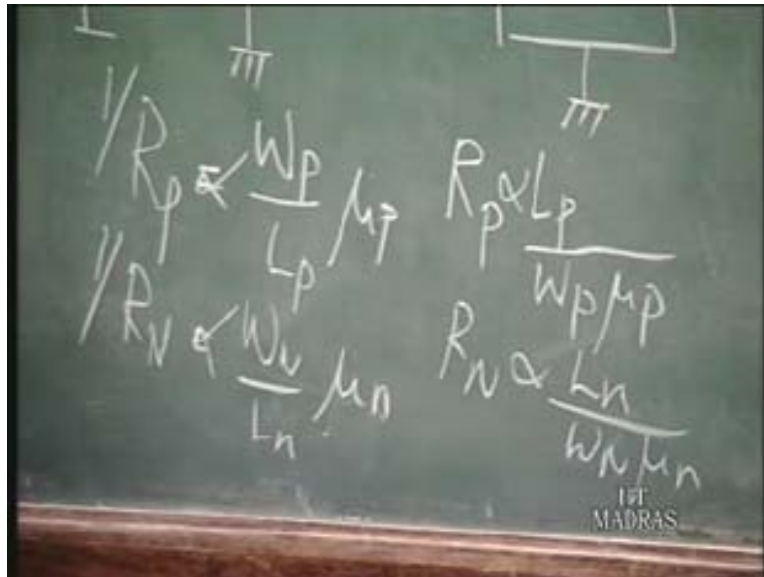
So if you have both the inputs low which means that both the NMOS devices are off and both the PMOS devices are conducting. So here both the PMOS device are conducting means that the output is connected to V_{DD} and the output is going to be high. If one of the input is high and the other low, it means one of this NMOS

transistors is conducting but they are in series which means that the lower path is off but the upper path is conducting because they are in parallel. So the output is connected to V_{DD} , output is one. What about 1 1? That is both the inputs are high so both the NMOS transistors are on, both the PMOS transistors are off. The output is connected to ground, so the output is low. This is the truth table of a NAND gate.

These are the two circuits so this is the circuit of a NOR gate where you have the two PMOS transistors in series, two NMOS transistors in parallel. This is the circuit of the NAND gate where you have the two NMOS transistors in series and the two PMOS transistors in parallel. So now there is a question. Now if I had the option to choose between these two circuits to realize any logic function which would I take? The NAND is more suitable because see R_{p1} and R_{p2} , the resistance of a p channel MOSFET is proportional to W_p by L_p into K_p whereas the resistance of the N channel MOSFET is W_n by L_n into K_n . W_p by L_p into μ_p and W_n by L_n into μ_n , so now μ_n is more than μ_p , so it is 1 by R_p , 1 by R_n .

So the resistance is proportional to, R_p is proportional to L_p by W_p μ_p and R_n is proportion to L_n by W_n μ_n .

(Refer Slide Time: 224:54)



So if you look at the NOR configuration where the two PMOS transistors are in series and the two NMOS transistors are in parallel. Here obviously the resistance of the NMOS transistors is less compared to that of the PMOS transistors because the mobility of the electrons is more than that of the holes by two and half times. Now suppose you choose the minimum size devices for all these devices that is if x is the minimum feature size, if all the devices are having the lengths and widths equal to x then what happens? Then in the NOR configuration the total resistance of the upper branch is

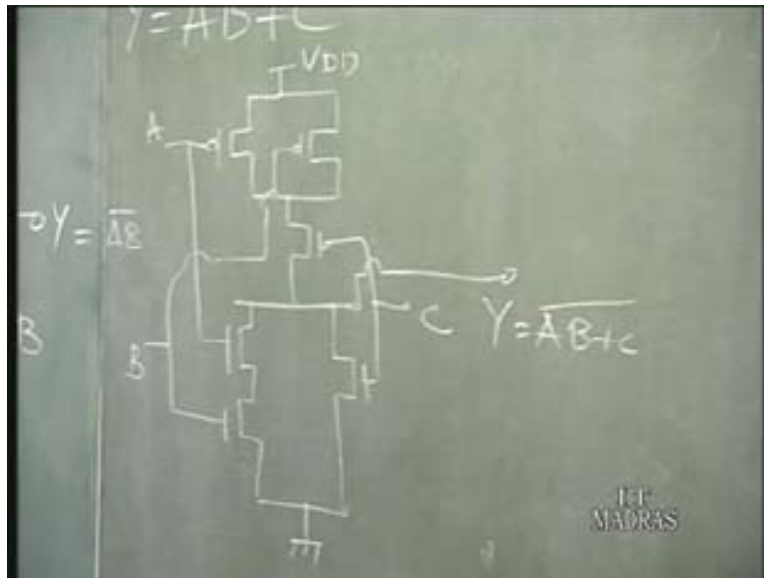
going to be is going to be say if you assume that R_p is greater than R_n by say around two and half times which is the ratio of the mobility in this case. So this is going to be five times the resistance of an N channel MOSFET whereas this is going to be half the resistance of an N channel MOSFET because they are in parallel. If you are charging or discharging a capacitance, the charging time is going to be about ten times more than that of the discharging time, if you use all minimum size devices.

Whereas if you go to the other configuration that is the NAND configuration now here the resistance of the p channel devices say is again two and a half times that of an NMOS device. So the resistance of this parallel branch is going to be around one and a quarter times of NMOS device whereas this one is this around twice. So you see that here there is not much difference between the two. So what is the message here that if you make a NAND gate, you can jolly well go for minimum size devices, all the transistors can be made having the minimum feature size that is widths and lengths equal to the minimum feature size.

Even then the outputs is going to be switching from the low to high and high to low almost equal times. So it's called equal drive capability. Basically you have almost equal drive capability for the low to high and the high to low transition. Whereas in this case there is a lot of asymmetric in the drive capability, when the output is going from high to low and low to high. So that way the NAND is going to be preferred because you can have minimum size devices which is going to reduce the total area required to fabricate the NAND gates. So that is why one would prefer the NAND gate.

Now of course these are the NAND gates and the NOR gates. Now if we want to achieve any other logic using this, you have to say for example let us take an example suppose you want to make $A B$ plus C bar then what you do? Here you have y is equal to AB bar. Now you have to modify this circuit. What do you do? You have another NMOS transistor in parallel with this series of A and B and you must have another PMOS transistor in series with the parallel combination of A and B . So I will draw the circuit, so this two transistors PMOS transistor, you have another transistor in series PMOS transistor and here you have the two NMOS transistors in series and you must have another NMOS transistor in parallel. So this goes to ground, this is VDD . So this goes here you may connect this here, you may connect this here. So you can have this input A , this input B and this input C and the output is from here. So this circuit gives you this logic function Y is equal to AB plus C bar.

(Refer Slide Time: 30:59)

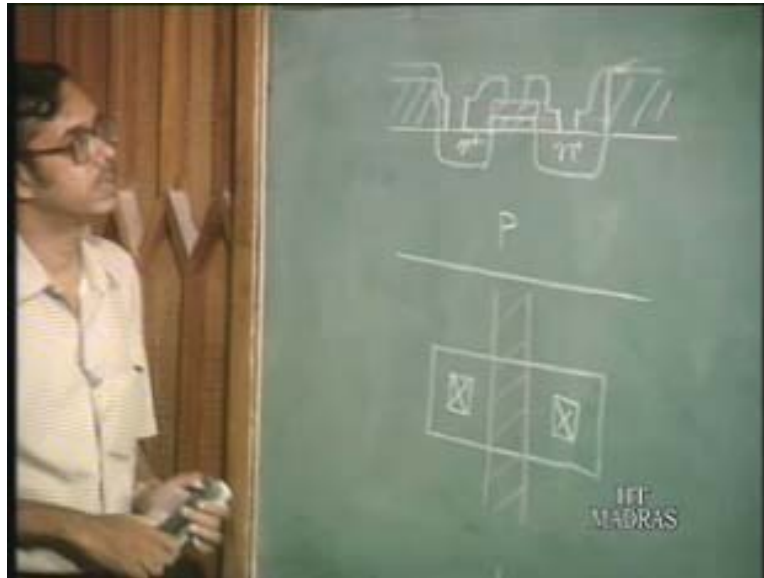


So the basic idea is whenever you have a series combination for the NMOS, you must have a parallel combination for the PMOS and when something is in parallel of NMOS configuration, it must come in series for the PMOS. So here one thing is compared to the NMOS logic circuits which we have gone through, here in a CMOS the number of transistors required is much more that is if you have a three input gate here, you require six transistors whereas in other NMOS logic circuits which we have seen you have only one load transistor and you will require three transistors in the driver part that is a lower half of the circuit. So of course this requires more number of transistors so that is one disadvantage. From CMOS point of view we shall look at other circuits later which remove this problem. I shall now digress a little bit and talk a little bit about the technology of fabricating MOS devices here. The MOS devices which are fabricated, the steps are very simple. So what you do is suppose you want to make an NMOS, you start with a P type substrate, you grow an oxide and then you define the active area by etching a part of the oxide. So you remove the oxide from a certain region where you are going to make the MOS device. So the MOS device it will be something like this from where the oxide has been removed. The next thing what is done is you grow a thin oxide which is going to be the gate oxide, so here you have the gate oxide grown. Here also the slight growth not much then you have this gate oxide, you don't require any mask for it. Then what you do is you deposit poly silicon which is the gate material on top of the oxide.

Again no mask is required here. Now you require a mask to define the gate area where you want the gate to be. Basically you have a mask something like this say, you have deposited the poly silicon. Now the mask for the gate is something like this which will leave the poly silicon in this region and remove the poly silicon in all other regions. So if you look here basically what you do is you remove the poly silicon in all regions and leave the poly silicon here and then using this poly silicon as mask you also remove the thin oxide in this regions so you are left with structure like this. Now what you do is

you subject, you do a diffusion N type diffusion. So what happens is all the exposed silicon regions get diffused with N type impurities. So you have n plus here which act as the source and drain and you also simultaneously dope the poly silicon.

(Refer Slide Time: 36:32)

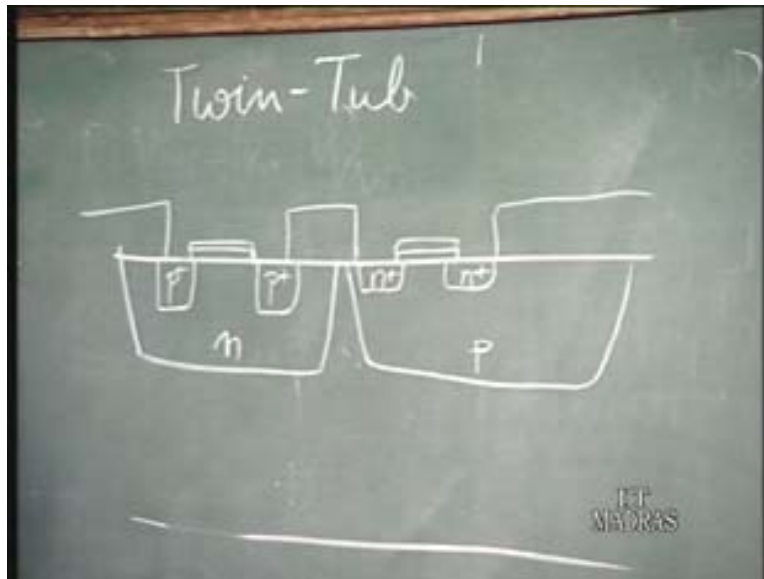


So this poly silicon also becomes n plus. Here also you don't require a mask as such. So basically if you look from the top this regions becomes n plus and this poly silicon also gets doped, only the channel region under the gate is not n plus.

So this is a self-align technology for making n mask. Later what you can do is to take the interconnections, you again grow oxide on top of this and then you take the contacts. Open a contact window here, you take the source and drain contacts and also if you want to take contact from the poly silicon gate you can open a contact here. So you have contact pairs here, you have another contact pair here and of course poly silicon, you may take a contact here you may not because poly silicon itself acts as a line and it can be used for interconnections itself.

So this is the structure of a MOSFET and of course then this may be followed by metallization where you deposit metal and take contacts and also may be for interconnections. So this is how you make a NMOS. For PMOS device the process would be identical except that the starting wafer should be N type and the diffusion which you are doing is P type for source and drain. Now in a CMOS you require to make both NMOS and PMOS. So how do you do it? Initially what was done is you start with a n type substrate, make what is called a P well like this and you make the PMOS devices here. So you have the PMOS devices here and you have NMOS devices in the P well say.

(Refer Slide Time: 38:48)



So this is the PMOS devices and this is the NMOS devices but the problem here is that the p type doping has to be more than the n type doping. If you want to make a P well in an n type wafer and the threshold voltages are dependent on the substrate doping concentration.

So it is difficult to get equal threshold voltages for the PMOS and NMOS devices. So what you have is presently the more popular is the twin top technology where you actually start off with a very lightly doped substrate and you actually make two tubs, one is an n tub where you make the PMOS devices and a p tub where you make the NMOS devices and the devices are fabricated in a similar way as we have fabricated the discrete NMOS device. Basically what you have here is if you go back to mask diagram here, the mask is very simple. You have a region where you have the diffusion done and you have another region where the gate is there and whenever these two regions overlap, you have a transistor. So basically this can be represented usually by what is called a stick diagram where you have instead of diagram like this, use lines to represent the poly silicon going like this and the diffusion is like this.

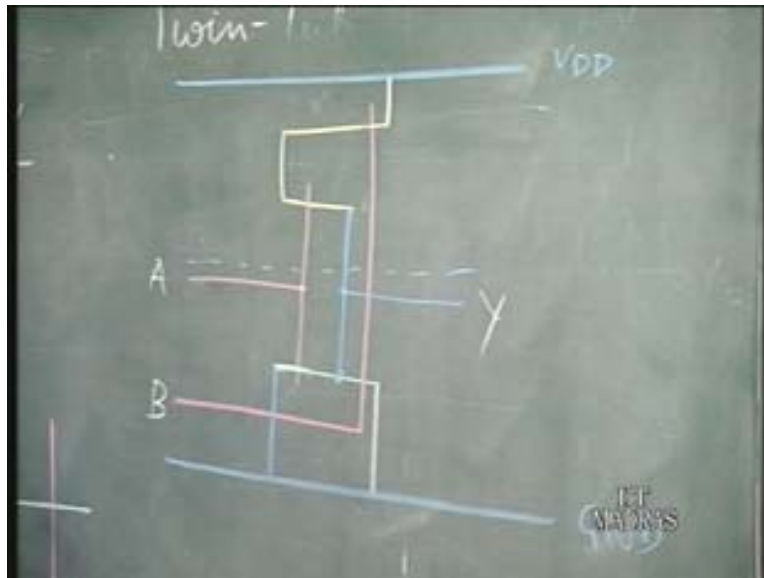
So wherever the diffusion and the poly silicon line overlap the diffusion, you have a transistor. The other lines can overlap without any problem that is the metal line can go over the poly silicon because usually the poly silicon will be covered by oxide, so there is no problem and also the metal can go over diffusion without any problem. So only when the poly silicon and diffusion overlap you have a transistor.

So for example if where to draw the stick diagram so you know this layout can be easily drawn with the stick diagram. For example if you take the NOR transistor, the NOR gate where you have the two NMOS transistors in parallel and the PMOS

transistors in series. So basically what you have? you will have a demarcation line to distinguish the n well and the P well and in the p well you will be fabricating the NMOS transistors and in the n well you will be fabricating PMOS transistors. So you have two lines here for the two metal lines, one is for the V_{DD} . so this is for V_{DD} . and this is the ground and then you have the diffusions.

So for the PMOS devices you have p diffusion so suppose I draw it like this and then for the NMOS devices I have the n diffusion which I draw with this color. Now the poly gate, so one of the gate contacts it goes like this. So this is one input, the other one goes like this and it cuts here. So now you have the two pairs of transistors here, the NMOS transistors they must be in parallel. So this line is shortage here and this is a metal line which comes here. So what you have now is this is one NMOS transistor and this is one NMOS transistor, the two NMOS transistors where the diffusion is overlapping the poly silicon and you see that the drains of the two NMOS transistors are shorted here and the sources are shorted here. So basically these two NMOS transistors are in parallel whereas here they are in series. This is one NMOS transistor and this is another NMOS transistor and they are in series. These are the metal lines. The output can be taken from here. The two inputs are can be taken through this poly silicon lines.

(Refer Slide Time: 44:44)



So this is one input here which goes to the gate of one PMOS and one NMOS here. Another input which goes to the gate one NMOS and one PMOS transistor. So this can be one input A, this can be one input B, this is the output Y. So this diagram gives you the idea of the layout or how you place the different transistors while fabricating the CMOS NOR gate. That is the diffusion say for example the p diffusion is going to go like this, the n diffusion goes like this, the poly lines go like this. So you take the inputs, the inputs are given through the poly lines and the output is taken here. So this is the metal line which interconnects the drain of this PMOS with the drain of this NMOS here. So this is the stick diagram which is used extensively to show the layout of this CMOS gates. You need not draw all of them as they are but just as lines.

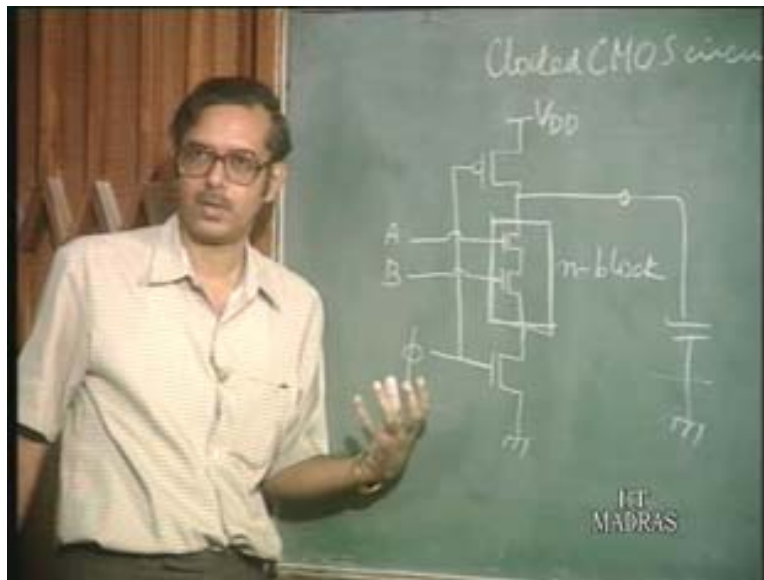
Similarly one can also draw the stick diagram for NAND gate where the PMOS transistors are going to be in parallel and the NMOS transistors are going to be in series. I think one can draw that quite easily. So this is to give you an idea of how to represent the layouts of the CMOS gates. So we have seen that the CMOS gates although they have their advantages of low power dissipation and because both the transistors, the upper half and the lower half are not conducting simultaneously but they have a disadvantage that the number of transistors required is quite large because with every input you have to have two transistors. So how can one reduce that without losing the advantage of a CMOS? So that can be done by using what are called dynamic or clocked CMOS circuits.

So we shall take that up next that is clocked CMOS circuits. Now you know that in a logic circuit, in a complicated logics circuit it will consist of sequential logic circuits as well as combinational logic circuits. Sequential logic circuit means flip flops which is an integral part of a system. So basically the clock is there, it's not that you have to make it and all these transitions have to be synchronized, even for a combinational

logic, these transitions have to be synchronized with respect to a clock.

So you can use the same clock to actually realize combinational logic even NAND and NOR for example or more complicated logic functions and with the result we can have circuits which require much less transistors compared to the CMOS circuits. Say for example let us just take the simple example of the NAND gate. So suppose I have of course in the NAND gate, it is just an example. So this is a block say, we call this n block because it consist of only NMOS transistors so this is VDD, this is the output. Suppose you have load capacitance and these are the two inputs A and B and here you apply the clock.

(Refer Slide Time: 49:51)



So when the clock goes low, what is happening is this NMOS transistor is going to be off and this PMOS transistor is going to be on. What is going to happen to the capacitance? The capacitance is going to be charged so this period is called the pre charge phase and then what happens? When the clock goes high, this PMOS is off, this NMOS is on. So now provided that this n block which is there it conducts. This capacitance is going to get discharged. that is if both the NMOS transistors are on then only the capacitance is going to get discharged that is if the two inputs are high then only the output is low. Otherwise the output is going to be high, just like in a NAND gate only if both the inputs are high then only the output is low.

So there are two phases here, one is called the pre charge phase where you charge the output capacitance and then the evaluate phase where you actually evaluate the logic function, when the output is actually a function of inputs. So this is just an example of a dynamic logic circuit or a clocked logic circuit, here you see that basically what you require is if for an n input circuit, you require n transistors in the n block plus 2, 1 NMOS and plus 1 PMOS.

So you require $n + 2$ transistors compared to $2n$ for a CMOS and otherwise it is just like a CMOS, you see what is happening is with every clock pulse, the output is getting charged in one phase and then its getting discharged, just like a CMOS. So the power dissipation is going to be exactly the same as that of a CMOS. At the same time you actually save the area of the circuit which is going to be much less compared to a CMOS. So we shall discuss this in more detail in the next class.